CS 134: Networks
Prof. Yaron Singer Spring 2017

1 Overview

Last lecture, we learned about the history of the web and some primary motivations for ranking Web pages. In the Information Age, designing good algorithms for information retrieval is particularly difficult due to the abundance of seemingly-relevant documents given a query. An understanding of the fundamental network structure of Web pages is crucial for addressing ranking questions, as we now discuss. In particular, we will analyze a ranking algorithm, a precursor to PageRank, known as Hyperlink-Induced Topic Search (HITS).

2 Hubs and Authorities

Given a sample of "reference" pages containing links to "result" pages that are relevant to a certain query, we can determine the importance of each of the result pages by counting the number of in-links it has from the reference pages (this technique is known as **Voting By In-Links**). Furthermore, we can then determine a *reference* page's value as equal to the sum of the votes received by all pages for which it voted. Figure 1 demonstrates both of these techniques for a given network. We can then use these to get still more refined values for the quality of the result pages, and so on, potentially forever. This repeated refinement technique is known as the **Principle of Repeated Improvement**, exemplified in Figure 2.

Why does repeatedly updating page values this way produce beneficial results? In fact, you can recognize the intuition behind this re-weighting of votes in the way we evaluate endorsements in our everyday lives. Suppose you move to a new town and hear restaurant recommendations from a lot of people. After discovering that certain restaurants get mentioned by a lot of people, you realize that certain *people* in fact had mentioned most of these highly recommended restaurants when you asked them. These people play the role of the high-value reference pages on the Web, and its only natural to go back and take more seriously the more obscure restaurants that they recommended, because now you particularly trust their judgment. This last step is exactly what we are doing in re-weighting the votes for Web pages.

More formally, the results pages we were originally seeking - the prominent, highly endorsed answers to the queries - are called the **authorities** for the query. The reference pages are called the **hubs** for the query. We want to find the value of each page p in a network as a potential authority and as a potential hub, which we refer to as a_p and h_p respectively. To do this, we can utilize the following definitions and algorithm, based on the Principle of Repeated Improvement:

Definition. ***Authority Update Rule****: For each page p, update a_p to be the sum of the hub scores of all pages that point to it.*

Definition. ***Hub Update Rule****: For each page p, update h_p to be the sum of the authority scores of all pages that it points to.*

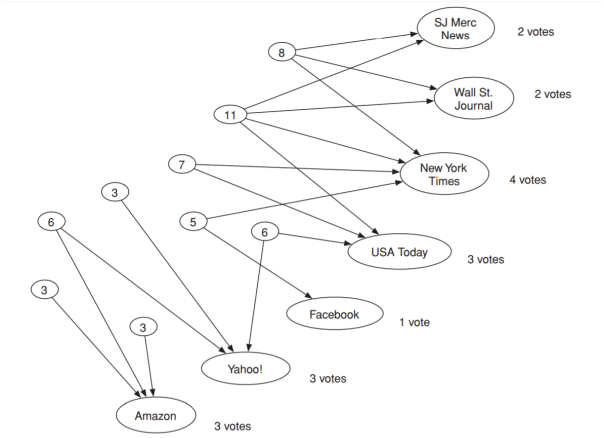

Figure 1: An example of Voting by In-Links. Note that each reference pages value is written as a number inside it (*Kleinberg 14.2*).

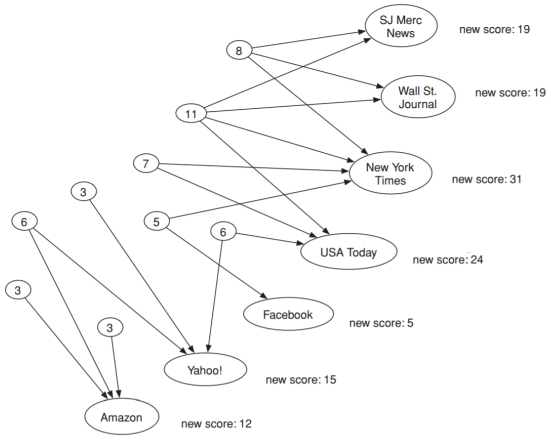

Figure 2: An example of the Principle of Repeated Improvement in action. Each result pages new score is equal to the sum of the values of all lists that point to it (*Kleinberg 14.2*).

ALG 1: HITS Algorithm

input: Graph G with n nodes, number of steps t

1. Start with all hub scores and all authority scores of each page equal to 1.
2. For t iterations:
   First apply the Authority Update Rule to the current set of scores;
   Then apply the Hub Update Rule to the resulting set of scores;
   End for
3. Normalize by dividing each authority score by the sum of all authority scores.
4. Normalize by dividing each hub score by the sum of all hub scores.

return: The hub and authority scores of each page

Notice how a single application of the Authority Update Rule (starting from a setting in which all scores are initially 1) is simply the original casting of votes by in-links. A single application of the Authority Update Rule followed by a single application of the Hub Update Rule produces the results of the original list-finding technique. In general, the principle of repeated improvement says that, to obtain better estimates, we should simply apply these rules in alternating fashion, giving us the Hyperlink-Induced Topic Search, or HITS, algorithm.

Just as we saw with PageRank, we would expect that the normalized values converge to limits as t goes to infinity, as they do for our in-links example in Figure 3. In fact, except for a few rare cases (characterized by a certain kind of degenerate property of the link structure), we reach the same limiting values no matter what we choose as the *initial* hub and authority values, provided only that all of them are positive. In other words, the limiting hub and authority values are purely a property of the link structure, not of the initial estimates we use to start the process of computing them. We will go on to prove this in the next section.

Ultimately, what these limiting values correspond to is a kind of equilibrium: their relative sizes remain unchanged if we apply the Authority Update Rule or the Hub Update Rule. As such, they reflect the balance between hubs and authorities that provided the initial intuition for them: your authority score is proportional to the hub scores of the pages that point to you, and your hub score is proportional to the authority scores of the pages you point to.

3 Spectral Analysis of Hubs and Authorities

Thinking about hubs and authorities in a network, we define an **adjacency matrix** M as follows:

$$M_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Figure 4 shows an example of an adjacency matrix for a given network. Observe from the Hub Update Rule that the hub score of node j is

$$h_j = M_{j1}a_1 + M_{j2}a_2 + \cdots + M_{jn}a_n.$$

Along similar lines, from the Authority Update Rule the authority score of node i is

$$a_i = M_{1i}h_1 + M_{2i}h_2 + \cdots + M_{ni}h_n.$$

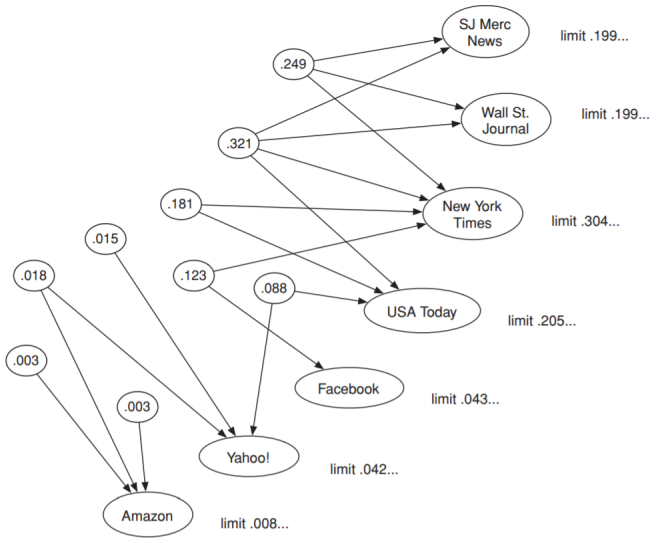

Figure 3: The limit hub and authority values for our in-links example (*Kleinberg 14.2*).

Then, letting $h^{(t)}$ and $a^{(t)}$ be the vectors of all hub and authority scores respectively at time step t, following the HITS algorithm we have:

$$a^{(t)} = M^T h^{(t-1)}$$

and

$$h^{(t)} = M a^{(t)}$$

where M^T is the matrix transpose of M. Observe that for the first few steps of the HITS algorithm, we get:

$$a^{(1)} = M^T h^{(0)}$$
$$h^{(1)} = M a^{(1)}$$
$$a^{(2)} = M^T h^{(1)} = M^T M M^T h^{(0)}$$
$$h^{(2)} = (MM^T)^2 h^{(0)}$$

More generally, we infer:

$$a^{(t)} = (M^T M)^{t-1} M^T h^{(0)}$$
$$h^{(t)} = (MM^T)^t h^{(0)}$$

Figure 5 shows how we might use our adjacency matrix to calculate a set of hub scores from a set of authority scores as we have determined above. This provides the groundwork for us to demonstrate the convergence of the scores. In other words, we first wish to show that there exists $h^{(*)}$, such that:

$$\lim_{t \to \infty} \frac{h^{(t)}}{c^t} = h^{(*)},$$

where c is a constant. Using our above equations we have:

$$(MM^T) h^{(*)} = c \cdot h^{(*)}$$

The constant c accounts for the fact that at each step, we normalize all the values. So the above claim is saying that the hub scores converge to some vector $h^{(*)}$ and are normalized by a factor $1/c$ at each step. Using linear algebra, we can infer that $h^{(*)}$ and c are an eigenvector and a corresponding eigenvalue of the matrix MM^T. We now utilize the following:

Definition. *(Review) A matrix A is **symmetric** if $A^T = A$.*

Theorem. *If a matrix A is symmetric then there exists n eigenvectors of A that are orthonormal. In other words, these eigenvectors form a **basis** for A.*

We observe that MM^T is symmetric since $(MM^T)^T = (M^T)^T M^T = MM^T$. Therefore, there exist eigenvectors $z_1, \cdots, z_n$ of MM^T with corresponding eigenvalues $c_1, \cdots, c_n$ that form a basis for MM^T and such that for all $x \in \mathbb{R}^n$, there exist constants $p_1, \cdots, p_n \in \mathbb{R}$ such that

$$\begin{aligned} MM^T x &= MM^T(p_1 z_1 + \cdots + p_n z_n) \\ &= p_1 c_1 z_1 + \cdots + p_n c_n z_n \end{aligned}$$

where the last equality follows from the fact that each z_i is an eigenvector. What this says is that $z_1, z_2, \cdots, z_n$ is a very useful set of coordinate axes for representing *any* n-dimensional vector x: multiplication by MM^T consists simply of replacing each term $p_i z_i$ in the representation of x by $c_i p_i z_i$.

If we assume without loss of generality that $|c_1| \geq \cdots \geq |c_n|$ and $|c_1| > |c_2|$, similar to our approach above we can write $h^{(0)}$ as a linear combination of the eigenvectors for some set of constants $q_1, \cdots, q_n \in \mathbb{R}$:

$$h^{(0)} = q_1 z_1 + \cdots + q_n z_n$$

Thus, we have:

$$\begin{aligned} h^{(t)} &= (MM^T)^t h^{(0)} \\ h^{(t)} &= c_1^t q_1 z_1 + \cdots + c_n^t q_n z_n \\ \frac{h^{(t)}}{c_1^t} &= q_1 z_1 + \left(\frac{c_2}{c_1}\right)^t q_2 z_2 + \cdots + \left(\frac{c_n}{c_1}\right)^t q_n z_n. \end{aligned}$$

From our assumption that $|c_1| > |c_i|$ for all $i \geq 2$, $\frac{h^{(t)}}{c_1^t}$ thus converges to $h^{(*)} = q_1 z_1$ as t approaches infinity. Thus in this case, even if we go with a new choice for the starting vector $h^{(0)} = x$, we see that $h^{(*)}$ will *always* converge to a vector in the direction of z_1, the only difference being the coefficient q_1 used when writing $h^{(0)}$ as a linear combination of the eigenvectors. It is in this sense that the limiting hub weights are really a function of the network structure, and not the starting estimates. (The case where $|c_1| = |c_2|$ will be discussed in section.)

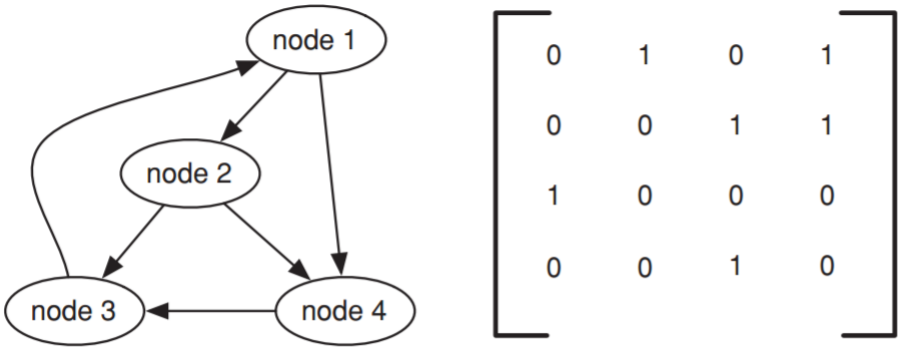

Figure 4: The directed hyperlinks among Web pages can be represented using an adjacency matrix M: the entry M_{ij} is equal to 1 if there is a link from node i to node j; otherwise, $M_{ij} = 0$ (*Kleinberg 14.6*).

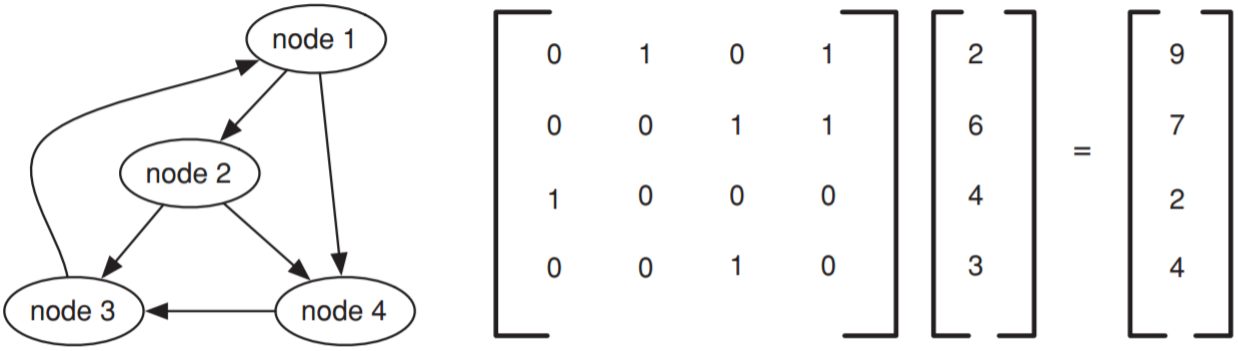

Figure 5: By representing the link structure using an adjacency matrix, the Hub and Authority Update Rules become matrix-vector multiplication. In this example, we show how multiplication by a vector of authority scores produces a new vector of hub scores (*Kleinberg 14.6*).